

Bradfordizing als alternativer Sacheinstieg: Evaluation thematischer Kernzonenbildung

Philipp Mayr

GESIS - Leibniz-Institut für Sozialwissenschaften, Bonn

33. Jahrestagung der Gesellschaft für Klassifikation (AG BIB)

18. März 2009, HTW Dresden

Agenda

- Einführung
 - Motivation
 - Fragestellung
- Bradfordizing-Studie
 - Thesen
 - Ergebnisse
 - Zusammenfassung
- Ausblick

Einleitung

Einordnung: Re-Ranking, Bradfordizing, Digitale Bibliothek

→ Schnittstelle aus Information Retrieval und Informetrie

Mehrwertverfahren für Informationssysteme zur
Verbesserung der Suche

Fokus: Re-Ranking durch Bradfordizing

→ Evaluation von Bradfordizing

Hintergrund I

Datenbank Perspektive:

- Große und heterogene Dokumentmengen zu fachspezifischen Fragestellungen
- Unterschiedliche relevante und zugängliche Datenbanken zu einem Thema (Sozialwiss., Politik, Bildung, Psychologie, Medizin ...)
- Fokus auf bibliographische Datenbanken (metadata)
 - Zeitschriftenartikel
 - Monographien

Hintergrund II

Nutzerperspektive (information seeking behavior):

- Relevante & qualitative Dokumente (relevance ranking)
- Umfassende Suche: Dokumente aus anderen Disziplinen (eine Datenbank ist nicht genug)
- Flexible Suchsysteme: alternative Suchstrategien und -techniken
- Mehrwerte, z.B. direkter Zugang zu Volltexten oder Metriken wie Zitationszahlen (Google Scholar)

Motivation - Idee

Anwendung und Evaluation des bibliometrischen
Verfahrens Bradfordizing (White, 1981) für das Information
Retrieval (IR)

... zur Erstellung und Analyse von Kernbereichen
(Sacheinstieg)

... fachbezogene Suchanfragen in bibliographischen
Datenbanken (Sozialwiss. und andere Disziplinen)

Bradfordizing zum Re-Ranking von Ergebnismengen

→ Informetrics und IR

Konkretisierung

- Große Ergebnismengen müssen nachträglich konzentriert werden (Kompensation, Strukturierung)
- Exploration eines alternativen Rankingprinzips zur Verbesserung der Informationssuche
- Hohe Plausibilität das der Kern einer Literatur einen Nutzen/Mehrwert in großen Ergebnismengen liefert
- Kontrollierte Vokabulare und darauf aufbauende Mehrwertdienste sind eine gute Basis um relevante Dokumente zu identifizieren

Bradford Law of Scattering

“If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the numbers of periodicals in the nucleus and succeeding zones will be as $a : n : n^2 : n^3 \dots$ “ (S.C. Bradford 1948)

- Beschreibung wie Artikel zu einem Thema (topic) über Zeitschriften verteilt (verstreut) sind
- Bradford hat die produktivsten Zeitschriften in einem Gebiet identifiziert
- Ziele:
 - um das ‘documentation chaos’ zu minimieren
 - um Kollektionen/Datenbanken zu analysieren und zu evaluieren
 - Später: als bibliometrische Untersuchungsmethode

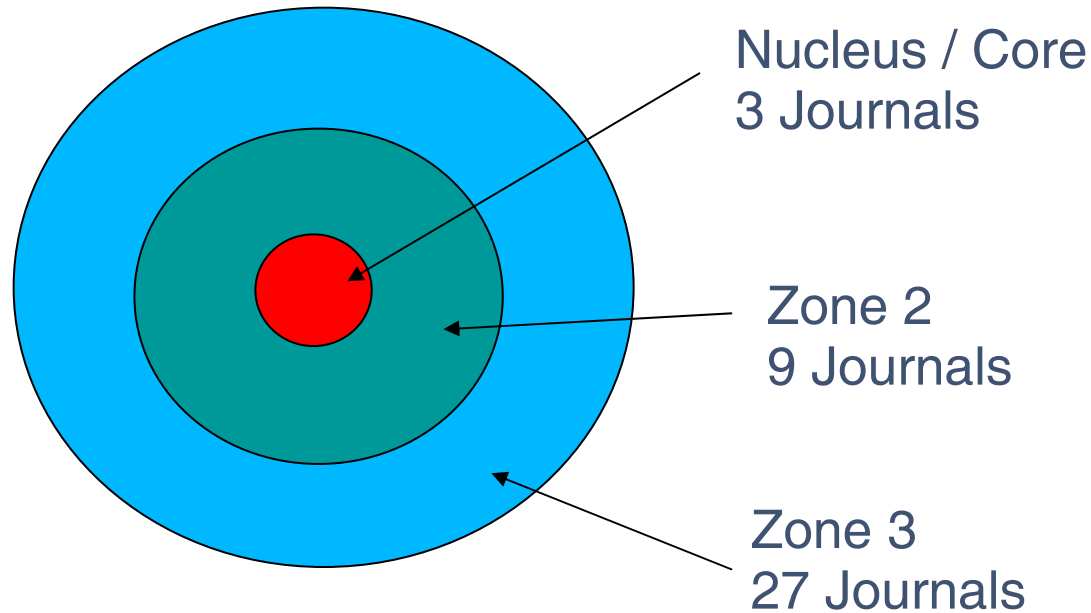
Bradford Law Beispiel

Idealized example: a literature (450 articles) for a subject e.g. „Systems theory“

150 Papers – 3 Journals (n)

150 Papers – 9 Journals (n²)

150 Papers – 27 Journals (n³)



Bradfordizing nach White

“... that is sorting hits

- (1) by the journal in which they appear, and then sorting these journals not alphabetically by title but
- (2) numerically, high to low, by number of hits each journal contains.

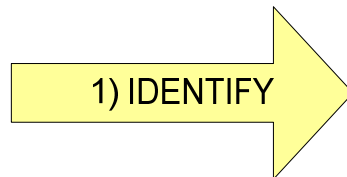
In effect, this two-step sorting ranks the search output in the classic Bradford manner, so that the most productive, in terms of its yield of hits, is placed first; the second-most productive journal is second; and so on, down through the last rank of journals yielding only one hit apiece.” (White, 1981)

→ Information streut auf eine charakteristische Weise

Ideales Szenario für Bradfordizing

Rang	ISSN	Artikel	Artikel kumuliert
1	0933-1883	3	3
2	0171-3957	2	5
3	0028-3355	1	6
4	0030-9273	1	7
5	0341-7069	1	8
6	0723-399X	1	9
7	0936-7780	1	10

Dokumentpool nach Bradfordizing

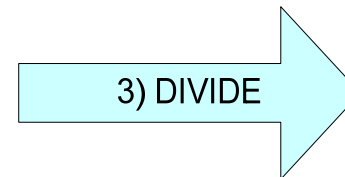


Controlled Vocabulary

Term Transformations



Bradfordized List



Present the nucleus for browsing

Bradfordizing-Studie

Fragestellung

Inwieweit lassen sich Bradford-Verteilungen in unterschiedlichen Fachgebieten und insbesondere bei fachlichen Recherchen in Literaturdatenbanken identifizieren?

Lassen sich informetrische Verfahren (z. B. Bradfordizing) als Mehrwertdienste für die Informationssuche in Literaturdatenbanken gewinnbringend nutzen?

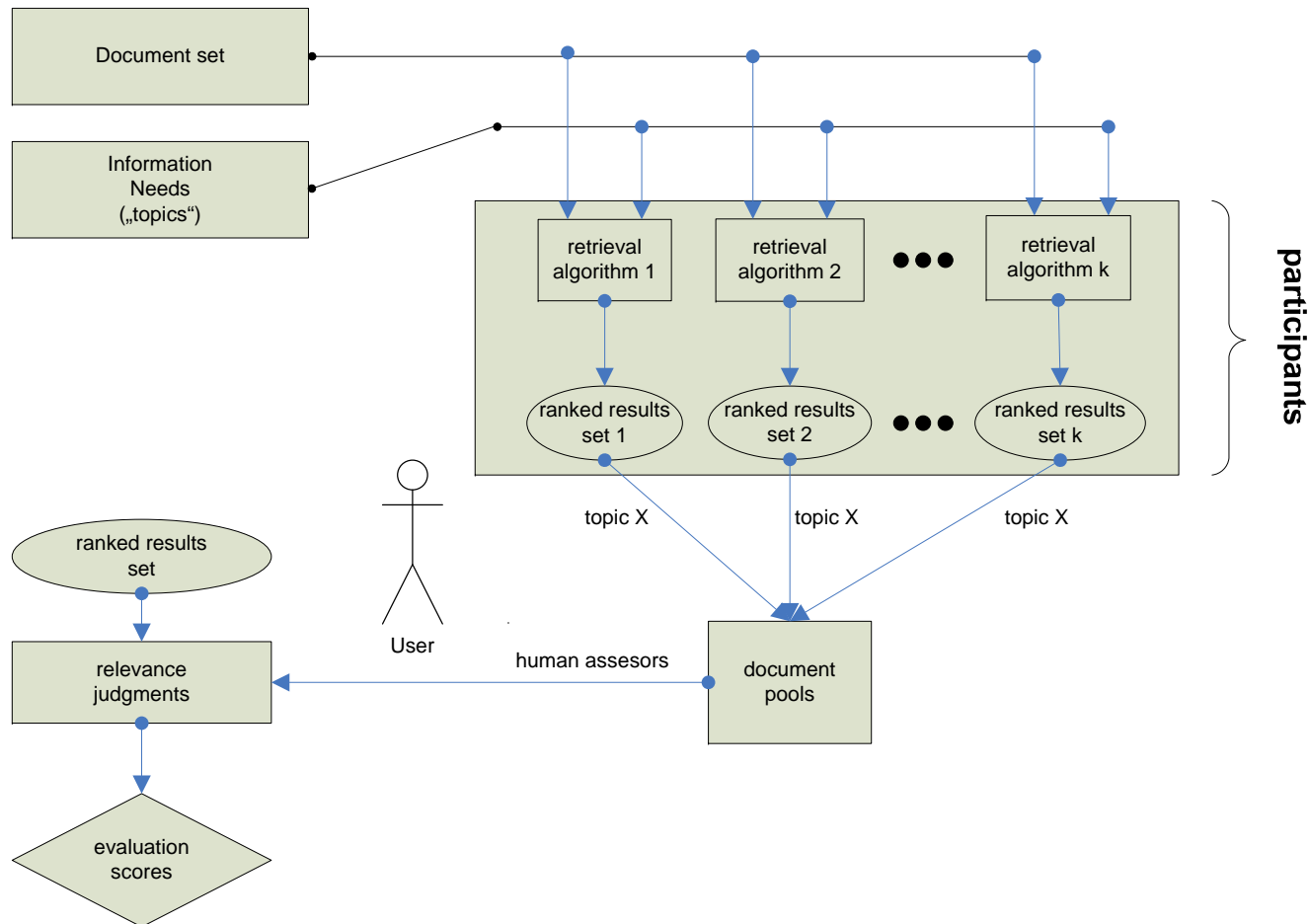
Welche Auswirkungen/Effekte hat die Anwendung von Bradfordizing als Re-Ranking-Ansatz für das Information Retrieval? Lassen sich die Effekte innerhalb eines IR-Tests evaluieren?

Lässt sich das Verfahren Bradfordizing als Re-Ranking-Ansatz auch auf andere Dokumenttypen und Szenarien übertragen? Kann Bradfordizing auch auf Monographien angewendet werden?

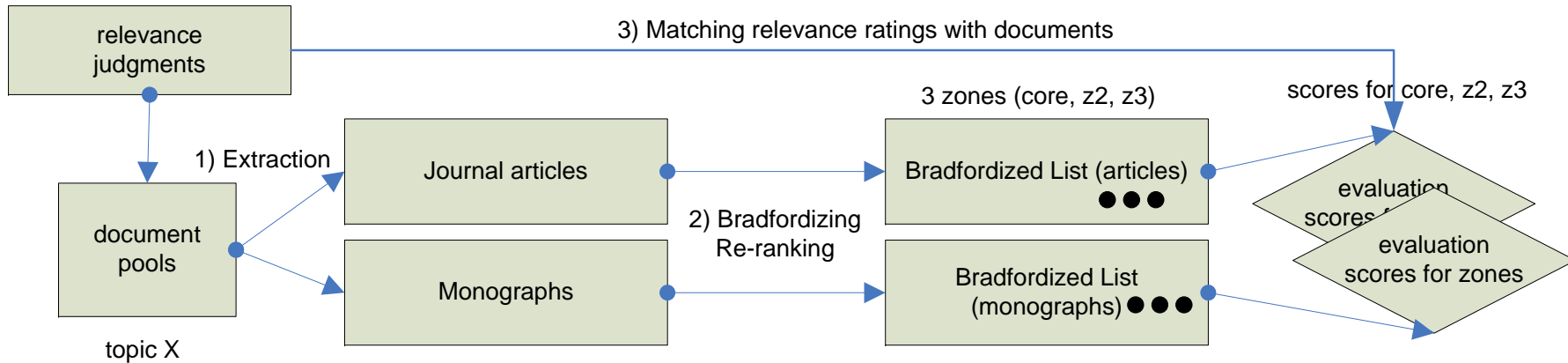
Methoden

- Bradfordizing (White, 1981) als Rankingverfahren
- Intellektuelle Relevanzbewertungen von Dokumenten (unterschiedliche Topics, Datenbanken)
- Interviews mit Informationprofessionals

Methode: klassisches IR Experiment

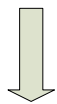


Methode: Evaluation von Bradfordizing



Core

Document 1
Document 2
Document 3
...
Document n



Relevant documents (P core)

Zone 2

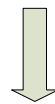
Document 1
Document 2
Document 3
...
Document n



Relevant documents (P z2)

Zone 3

Document 1
Document 2
Document 3
...
Document n



Relevant documents (P z3)

- Vergleich der Zonen
- Durchschn. Precision
- je Dokumenttyp

Methode: Beispielberechnung P und R

	Retrieved_article	Relevant	Precision _z	Recall _z
Σ Zone				
1 + Zone				
2 + Zone				
3	208	80	0,38	k. A.
Zone 1	73	41	0,56	0,51
Zone 2	65	25	0,38	0,31
Zone 3	70	14	0,20	0,18

Recall = relevante Dokumente n gefundene Dokumente / alle relevanten Dokumente

Precision = relevante Dokumente n gefundene Dokumente / alle gefundenen Dokumente

Topics, Dokumente und Datenbanken

	CLEF	KoMoHe
Zeitraum	2003-2007	2007
Topics	125	39
Documents total	65.297	31.155
Documents bradfordized	29.157	22.332
Journal article	18.112	17.432
Monographs	11.045	4.900
Databases involved	2 (1)	6

Topic im Standardformat

<top>

<num> 091 </num>

<DE-title> Qualität von Bildungssystemen </DE-title>

<DE-desc> Finde Dokumente, die die Sicherung der Qualität von Bildungssystemen in verschiedenen Staaten thematisieren. </DE-desc>

<DE-narr> Relevante Dokumente berichten über Modelle, politische Ziele oder theoretische Ansätze zur Sicherung der Qualität des Outputs von Bildungssystemen. Hierbei sind zum Beispiel Maßnahmen zum prozessualen Controlling oder Finanzcontrolling, Evaluationsmodelle oder Systeme der Netzwerkbildung zur Kooperation von an Bildung beteiligten Institutionen von Interesse. Wichtig ist, dass die Dokumente qualitative Aspekte der Bildung behandeln. </DE-narr>

</top>

Dokument im GIRT4-Format

```

43 <DOC>
44 <DOCID>iz-solis-90123690</DOCID>
45 <IDENTIFIER1>19900100587</IDENTIFIER1>
46 <ISSN>0044-3360</ISSN>
47 <TITLE-DE>Politik ohne Vertrauen?</TITLE-DE>
48 <DOCTYPE>journalarticle</DOCTYPE>
49 <SOURCE>Zeitschrift fuer Politik Organ der Hochschule fuer Politik Muenchen, N. F.</SOURCE>
50 <MEDIATYPE/>
51 <AUTHOR>Backes, Uwe</AUTHOR>
52 <PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
53 <LANGUAGE-CODE>de</LANGUAGE-CODE>
54 <CONTROLLED-TERM-DE>Bericht#Gesellschaft#Industriestaat#Inhalt#Konferenz#Osteuropa#
55 Politik#Politikwissenschaft#Vertrauen#westliche Welt#20. Jahrhundert</CONTROLLED-TERM-DE>
56 <ABSTRACT-DE>Der Beitrag berichtet ueber die Jahrestagung der Deutschen Gesellschaft fuer
57 Politikwissenschaft am 17. und 18.11.1989 in Erfstadt bei Koeln und fasst die wesentlichen
58 Inhalte der Referate zusammen. Leitthema war die Frage 'Politik ohne Vertrauen?', die sich zum
59 einen auf den diagnostizierten Vertrauensverlust demokratischer Institutionen und politischer
60 Parteien bezog, zum anderen durch die Wandlungsprozesse in den Staaten Osteuropas inspiriert
61 war. Die Tagung wurde eroeffnet mit Referaten zur 'grundsaeztlichen Bedeutung der Kategorien
62 'Vertrauen' und 'Misstrauen' fuer das Funktionieren politischer Systeme im allgemeinen,
63 konstitutioneller Demokratien im besonderen'. Es folgten Ausfuehrungen zum messbaren Ausmass
64 an Vertrauen bzw. Misstrauen der Bevoelkerung in westlichen Laendern gegeneueber ihren
65 polititischen Repraesentanten, zur Rolle des Vertrauens in der internationalen Politik, zu
66 Vertrauensbildung, Konfliktschaerfung und Friedensstiftung bei den Voelkern des Nahen
67 Ostens. Abschliessend wurden am Beispiel Polens drei Dimensionen des Vertrauens in den
68 Laendern Osteuropas untersucht: gemeinsame Werte als verbindlich geltende Ideologie,
69 politische Prozesse der Parteierrschaft und Effizienz der Parteierrschaft. Auf diese drei
70 Dimensionen erstreckte sich Vertrauen bzw. Misstrauen. (AG)</ABSTRACT-DE>
71 <TEXT/>
72 </DOC>

```

Relevance assessed documents from 6 databases

- 125 Topics from the CLEF-Evaluation (SOLIS)

- 39 Topics from the KoMoHe-Evaluation (SOLIS, SoLit, USB-Köln, Psyndex, WAO, Medline)

Ergebnisse

Ergebnis 1

Bradford-Verteilungen treten erfahrungsgemäß in allen Fachgebieten und damit auch bei Recherchen in Datenbanken auf.

Folglich kann das Bradfordizing zur Optimierung von Recherchen in der hier vorgeschlagenen Weise generell für themenspezifische Anfragen in Fachdatenbanken verwendet werden.

Ergebnis 1

CLEF-Jahrgang	Topics	Zeitschriftenartikel	Zeitschriften			
			total	core	z2	z3
2003	25	144	64	5,20	16,52	42,68
2004	25	115	60	5,28	17,76	36,52
2005	25	142	61	4,64	17,24	39,56
2006	25	153	75	7,08	20,76	47,40
2007	25	170	76	6,20	20,08	49,84

125 Topics des Projekts CLEF (GIRT-Korpus)

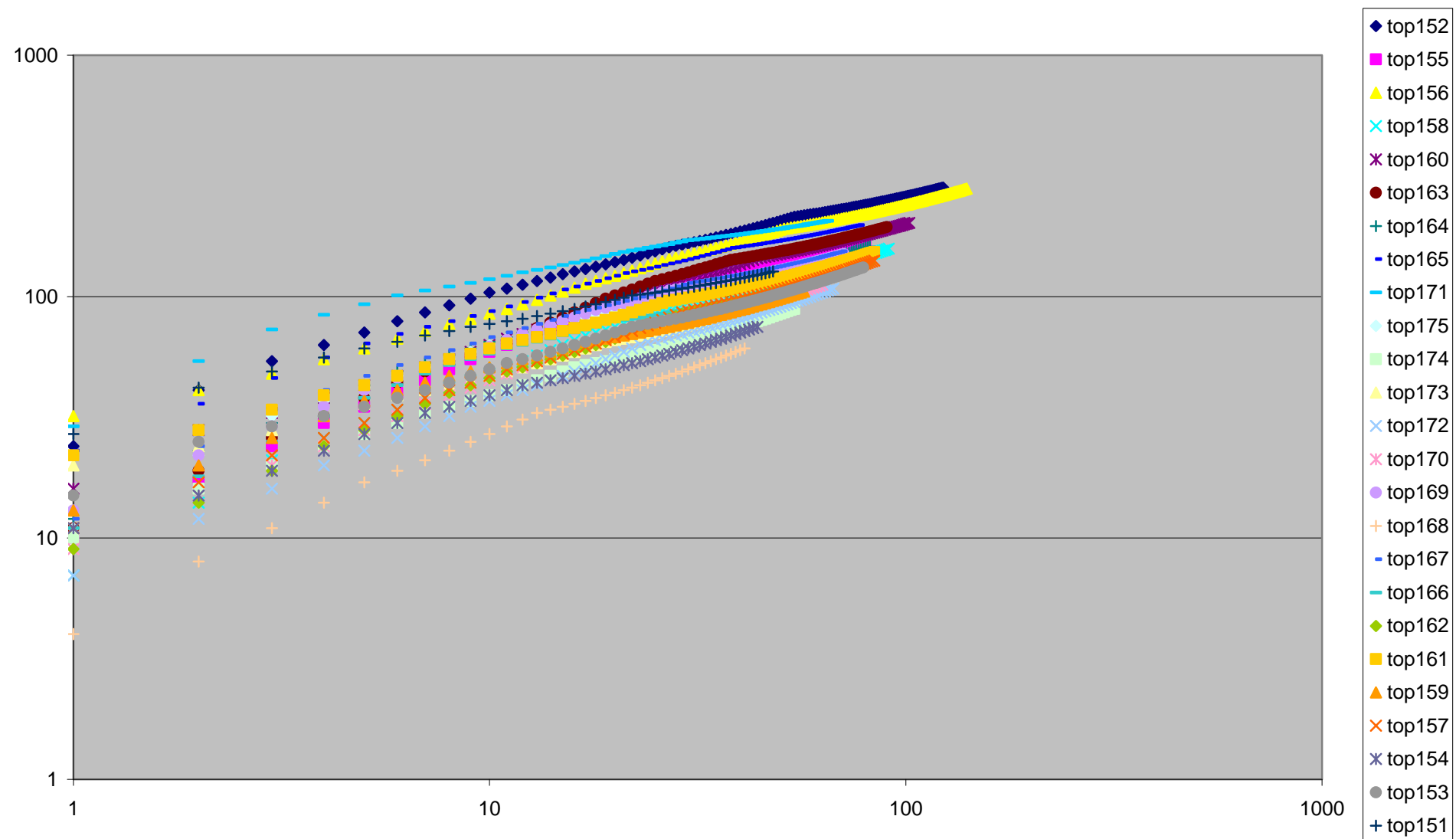
KoMoHe-Reihe	Topics	Zeitschriftenartikel	Zeitschriften			
			total	core	z2	z3
Test 1	12	449	134	8,25	27,25	98,67
Test 2	15	309	96	6,47	20,47	68,67
Test 3	12	688	250	15,92	54,83	179,33

39 Topics des Projekts KoMoHe (6 Datenbanken)

- Bradford-Verteilungen treten für alle Datenbanken auf
- trotz künstlicher Begrenzung des IR Tests insb. bei CLEF

Dokumentenverteilung: nach Bradfordizing

CLEF topics 2006



Ergebnis 2

Die Anwendung informetrischer Verfahren zum Re-Ranking von Dokumenten vermag eine alternative Sicht auf Ergebnismengen zu ergeben, die vom Benutzer intuitiv als positiv empfunden wird.

Positiv wird vor allem der Neuigkeits- und Erkenntnisgewinn bewertet, der durch hoch zitierte Arbeiten, Arbeiten zentraler Autoren, Artikel aus Kernzeitschriften und die Relevanzhierarchie des neuorganisierten Suchergebnisses entsteht.

Ergebnis 2

Antworten zu Frage 1 „spontane Nennung von Kernzeitschriften“	Topics
keine Nennung	13
1 Kernzeitschrift	4
2 Kernzeitschriften	3
mehr als 2 Kernzeitschriften	2
weitere Zeitschriften	8

Frage 1: Nennen Sie mir bitte spontan mehrere Zeitschriften, die Sie zu den Kernzeitschriften (häufigsten Zeitschriften) des Themas (XY) in der Datenbank SOLIS zählen.

N=24 Topics, qualitatives Interview

- Für mehr als 50 % der Topics können die „Informationsexperten“ „ihre“ Kernzeitschriften nicht spontan benennen.
- Mehrheit der Befragten attestiert dem Verfahren Bradfordizing (Kernzeitschriften als erstes ranken) einen positiven Relevanzeffekt.
- Größter Mehrwert entsteht vermutlich für unerfahrene Nutzer und Neulinge in einer Forschungsdisziplin.

Ergebnis 3

Die Anwendung des Bradfordizings bzw. das Re-Ranking nach Kernzeitschriften für thematische Dokumentmengen führt zu signifikanten Verbesserungen der Precision zwischen den drei Zonen (Core, Zone 2 und Zone 3).

Die Kernzeitschriften (Core) beinhalten signifikant mehr relevante Dokumente als Zeitschriften der Zone 2 oder der Zone 3.

Der größte Precision-Gewinn ergibt sich zwischen Core und Zone 3-Zeitschriften.

Ergebnis 3

CLEF article	Verbesserung core zu Z3 in %	Verbesserung core zu Z2 in %	Verbesserung Z2 zu Z3 in %	Verbesserung core zu base-line in %
2003	86,56 (*)	34,57 (*)	38,63 (*)	32,65 (*)
2004	69,23 (*)	22,45	38,20	26,25 (*)
2005	78,03 (*)	29,05 (*)	37,95 (*)	29,52 (*)
2006	17,63	7,66	9,27	8,46
2007	28,18 (*)	8,31	18,35	11,77
	55,93 (*)	20,41 (*)	28,48 (*)	21,73 (*)

Vergleich der durchschnittl. Precision zwischen den Zonen ergibt:

- Core relevanter als Zone 2 und Zone 3
- Zone 2 relevanter als Zone 3
- Meist signifikante Verbesserungen (t-Test, Wilcoxon)

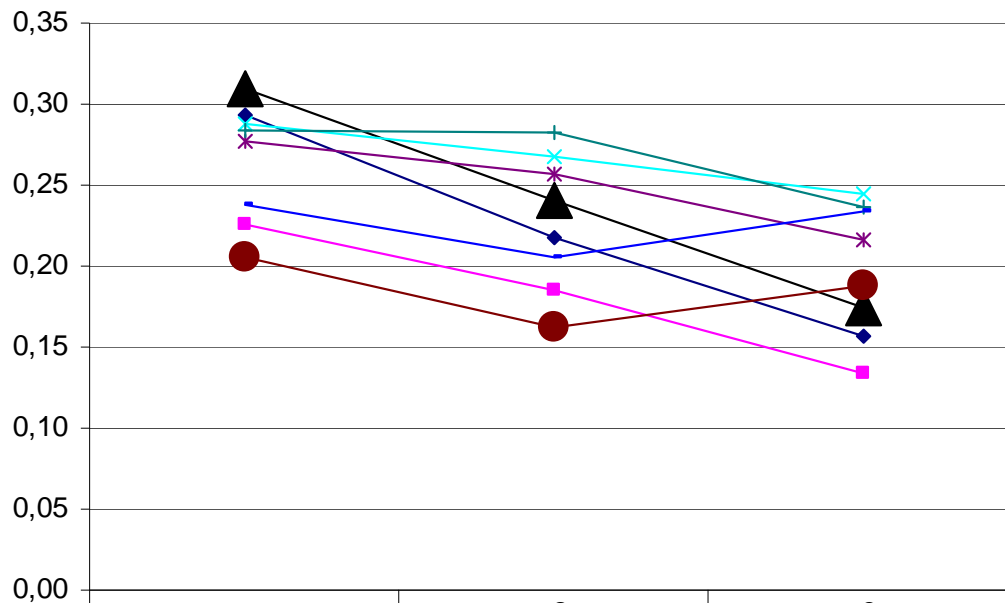
- niedrigere Verbesserungen bei KoMoHe
- kontinuierliche Verschiebung der Relevanz

→ relevance related distributions

KoMoHe article	Verbesserung core zu Z3 in %	Verbesserung core zu Z2 in %	Verbesserung Z2 zu Z3 in %	Verbesserung core zu base-line in %
Test1	18,82	11,75	6,32	9,84
Test2	11,58	6,16	5,11	6,12
Test3	19,32 (*)	8,67 (*)	9,80 (*)	9,00 (*)
	16,57 (*)	8,86	7,08 (*)	8,32 (*)

* Significant based on the Wilcoxon signed-rank test and the paired t-Test

Precision pro Zone (core, z3, baseline)



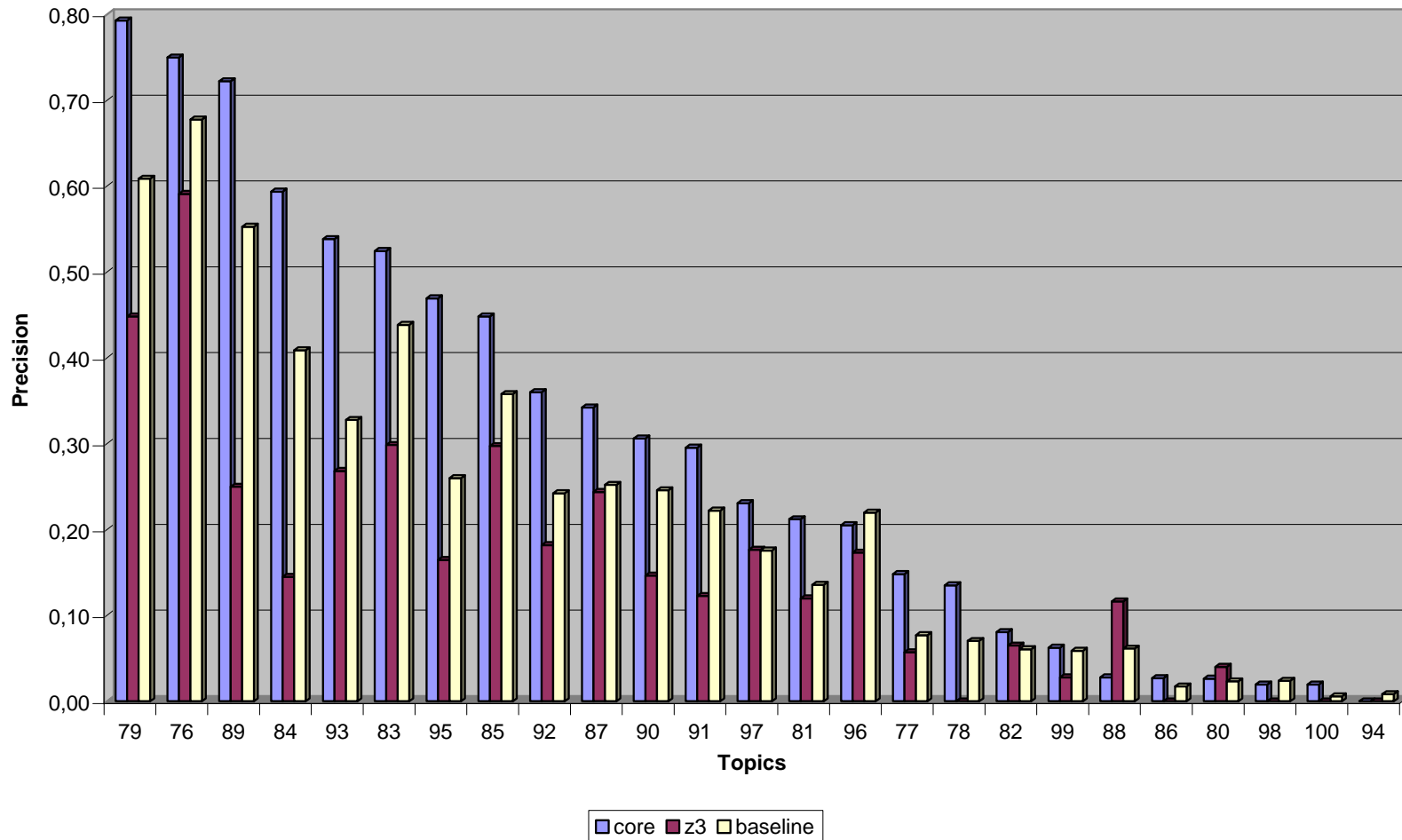
	core	z2	z3
◆ 2003 articles	0,29	0,22	0,16
■ 2004 articles	0,23	0,18	0,13
▲ 2005 articles	0,31	0,24	0,17
✕ 2006 articles	0,29	0,27	0,24
* 2007 articles	0,28	0,26	0,22
● 2005 monographs	0,21	0,16	0,19
+ 2006 monographs	0,28	0,28	0,24
■ 2007 monographs	0,24	0,21	0,23

Bradford zones (core, z2, z3)

Tab:
Precision (av) for
journal articles and
monographs in the
different zones
(core, z2, z3)

Precision pro Zone (core, z3, baseline)

CLEF 2003 articles



Example: improvement of precision

Articles for 25 topic (CLEF 2005)	Precision	Improvement
all articles (baseline)	0.239	
articles in core	0.310	(29.52%)*
articles in z3	0.174	
articles in core	0.310	(78.03%)*
Monographs for 25 topic (CLEF 2005)		
all monographs (baseline)	0.188	
monographs in core	0.205	(8.98%)
monographs in z3	0.188	
monographs in core	0.205	(9.09%)

* Significant based on the Wilcoxon signed-rank test and the paired T-Test

Ergebnis 4

Bradfordizing für thematisch konzentrierte Dokumentmengen lässt sich erfolgreich auf Monographien (Verlag als Selektionskriterium) übertragen.

Die Anwendung des Bradfordizings bzw. das Re-Ranking nach Verlagen für Monographien führt zu geringeren Verbesserungen der Precision zwischen den drei Zonen (Core, Zone 2 und Zone 3).

Der größte Precision-Gewinn ergibt sich auch hier wie bei den Zeitschriften zwischen Core und Zone 3-Verlagen.

Ergebnis 4

CLEF books	Verbesserung core zu Z3 in %	Verbesserung core zu Z2 in %	Verbesserung Z2 zu Z3 in %	Verbesserung core zu base-line in %
2005	9,09	26,69	-13,89	8,98
2006	19,98	0,33	19,58	13,07
2007	2,09	16,13	-12,09	5,21
	10,39	14,38	-2,13	9,09

- Verbesserung können für Monographien bestätigt werden
- Insgesamt geringere Verbesserungen (kaum signifikant)
- Artefakt: z.T. z3 besser als z2

KoMoHe books	Verbesserung core zu Z3 in %	Verbesserung core zu Z2 in %	Verbesserung Z2 zu Z3 in %	Verbesserung core zu base-line in %
Test1	31,51	36,27	-3,50	26,58
Test2	4,80	1,57	3,17	4,56
Test3	10,91	11,83	-0,82	7,41
	15,74 (*)	16,56	-0,38	12,85 (*)

* Significant based on the Wilcoxon signed-rank test and the paired t-Test

Ergebnis 5

Erstaunlicherweise zeigen die vorliegenden Ergebnisse, dass nach dem Bradfordizing Aufsätze in Kernzeitschriften häufiger als relevant für die Fragestellung bewertet werden, obwohl bei Bradford ja in allen Zonen nur relevante Dokumente Eingang finden.


Erklärungsansätze für den Relevanzvorteil der Kernzeitschriften sind:

- die Tatsache, dass Kernzeitschriften in der Regel den State of the Art eines Themas repräsentieren,
- die Tatsache, dass Aufsätze in Kernzeitschriften, die das sog. peer-reviewing durchlaufen, von Spezialisten dieses Faches bewertet werden und nicht von eher fachfremden und
- die Tatsache, dass bei den Publikationen im Kernbereich mehrere Aspekte des jeweiligen Forschungsgebietes abgedeckt werden.

Zusammenfassung

- Empirischer Nachweis, dass Bradfordizing konkrete Effekte für das IR generiert (164 Topics, 6 Datenbanken, 2 Dokumenttypen)
 - Nachweis der thematischen Relevanz von Kernzeitschriften
 - Zusammenwirken mit weiteren Mehrwertdiensten erscheint sehr plausibel und vielversprechend (insb. Crosskonkordanzen)
 - Konkretisierung von Bradford's Law
- Ranking auf Basis von szientometrischen Methoden

Ausblick

Benutzer sucht nach Themen 

2) Search Term Recommender (STR) schlägt geeignete kontrollierte Terme vor

3) Benutzer kann geeignete kontrollierte Terme auswählen

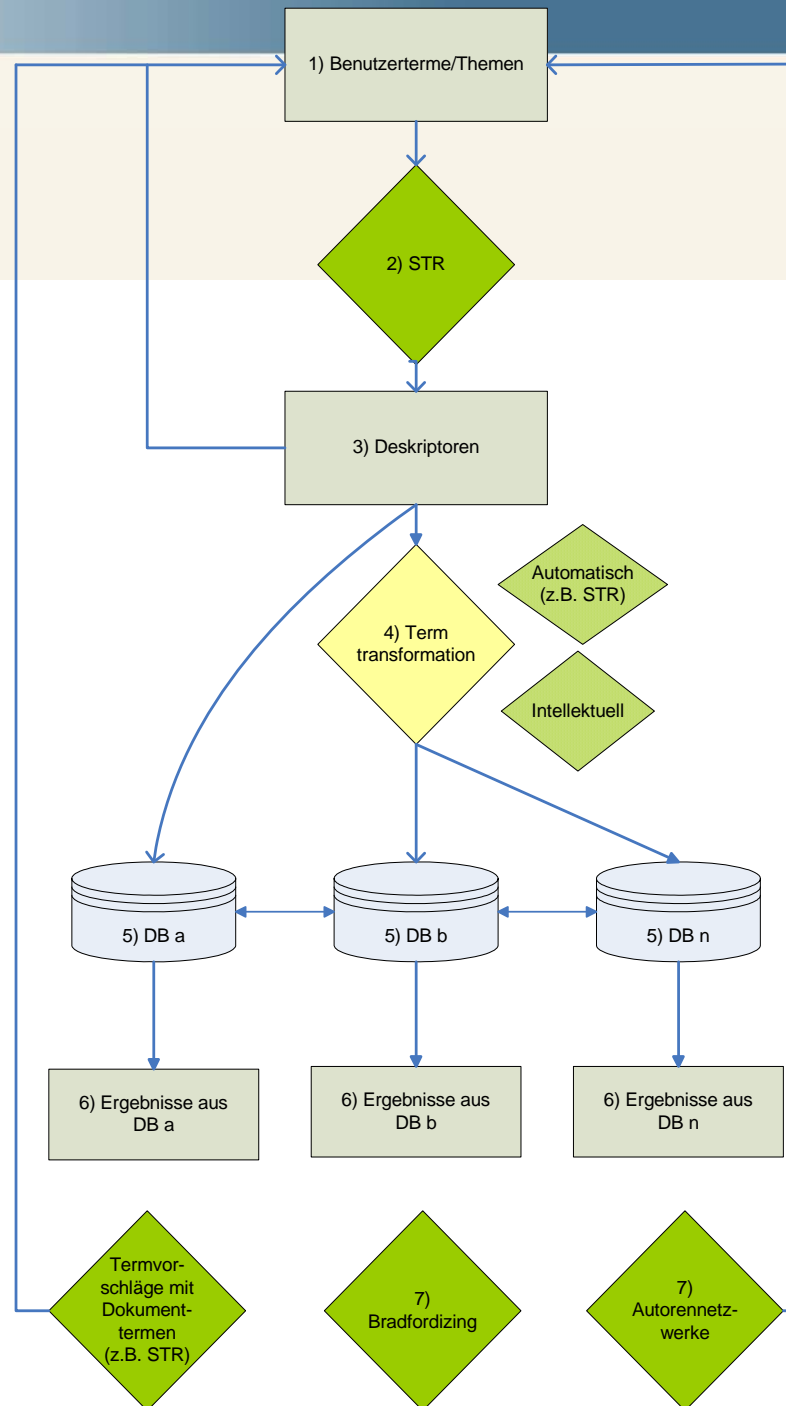
4) Ausgewählte Deskriptoren werden in verbundene Vokabulare transformiert

Datenbankspezifische Deskriptoranfragen werden automatisch generiert

5) Verteilte heterogene Datenbanken werden durchsucht

6) jede abgefragte Datenbank liefert ein Ergebnis

7) Re-Ranking der Ergebnismengen



Kombination unterschiedl. IR-Mehrwertdienste

- STR (2)
- Crosskonkordanzen (4)
- Re-Ranking (7)

Publications

Mayr, Philipp (2009): **Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in Digitalen Bibliotheken**. Philosophische Fakultät I, Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin. 237 p.,

URL: <http://edoc.hu-berlin.de/dissertationen/mayr-philipp-2009-02-18/PDF/mayr.pdf>

Mayr, Philipp; Mutschke, Peter; Petras, Vivien (2008): **Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking**. In: Library Review 57, No. 3, pp. 213-224.

URL: <http://arxiv.org/abs/0712.2449>

Danke für Ihre Aufmerksamkeit!

Kontakt:

Philipp Mayr
GESIS - Leibniz Institute for the Social Sciences

Lennéstr. 30, 53113 Bonn, Germany

<mailto:philipp.mayr@gesis.org>
<http://www.gesis.org/>